

Enceavra Research Round 41: Hard by Construction or Hard by Selection? Disagreement Mined Item Pools, Isomorphic Twin Holdouts, and the Consensus Errors That Disagreement Cannot Find

Date: 2026-10-30

Research lab: Enceavra Research

Short code: HPS-41 (Hard Pool Selection)

Primary tracks: Collaborative track ("When does dividing reasoning across specialists help, and when does it fragment evidence?"), Track 5 (Calibration, Abstention, and Release Control)

Position in the program: first round of the proposed Round 41 to 50 "Theory v1" block. Its proposed focus, from the Round 40 handoff, is a hard pool: items on which witnesses disagree, so that composition topologies separate on accuracy.

Carry-forward boundary: Round 40 (NCT-40) compared six composition topologies on 33 items and found that 32 of them had identical direct verdicts from all four witnesses, so accuracy could not rank topologies. Round 41 does not rerun that comparison. It asks whether selecting items by witness disagreement is a sound way to build the harder pool that Round 40 asked for.

Abstract

Round 40 ended with a request: build a pool of items on which the four witnesses disagree, so that topologies separate. The most direct way to do that is to screen many items and keep the ones that split the panel. That procedure has a known risk from benchmark construction: items chosen because a panel erred on them overstate how hard they are, and they understate the errors on which the whole panel agrees. Round 41 tests the procedure before using it. Each authored item is written twice as isomorphic twins with an identical rule trace and verdict: a screening twin is used to select, and the other twin is held out to evaluate.

We report three strictly separated evidence categories.

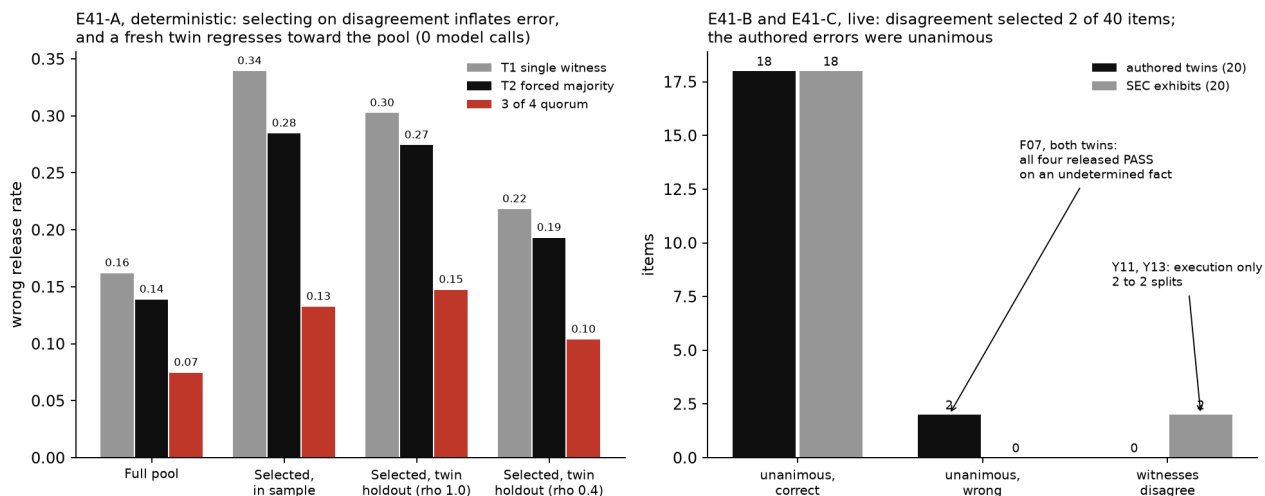
Deterministic (E41-A, no model calls). An exact enumeration over four witnesses, three verdicts, and three latent item types (easy, hard, and trap, where trap errors all fall on the same wrong answer) shows that disagreement selection keeps 42.2 percent of items and doubles the share of hard and trap items among them. On the selected items, the forced majority's wrong release rate is 0.285 in sample against 0.139 on the full pool, and the gap between forced majority and a single witness is more than twice as large in sample (0.055) as on the full pool (0.023). On a fresh isomorphic twin the gap falls back to 0.028. 13.4 percent of all forced majority errors sit on items that disagreement never selects, and a 3-of-4 quorum releases all of them.

Live authored (E41-B, 10 families, 20 twins). Ten families were authored to be hard by construction (holiday business day counts, an exact fifty percent waiver, a notwithstanding override, an incomplete cure, a missing schedule, and others). Screening produced 0 families with any disagreement, so the selection step selected nothing. All four witnesses answered 9 / 10 families correctly on both twins

and gave the same verdict on both twins of every family (40 / 40 witness twin pairs stable). The one error was unanimous: on F07, all four witnesses released PASS on both twins when the governing fact (whether later financial statements were delivered) was not in the packet. Only the disagreement routed hybrid (T6) held it, in exchange for one false hold.

Live real exhibits (E41-C, 20 SEC EDGAR exhibits fixed in advance). On twenty new real exhibits chosen for surface traps, eighteen produced unanimous correct decisions. Two produced the same 2-2 split, both in the execution only stratum: Y11 and Y13, amendments that become effective only on execution and delivery, or receipt, of counterparts by every party. On both, Witness A and Witness C certified and Witness B and Witness D held. A single witness and the forced majority (ties broken by Witness A) released both; the quorum and the extraction resolver held both.

Interpretation, kept separate. Disagreement is a detector of decorrelated error, not of difficulty. On this panel, the errors that remain are correlated, so mining for disagreement returns an empty or nearly empty pool and leaves the remaining errors exactly where it cannot look. The harder pool the Round 40 handoff asked for has to be built by construction, with labels fixed before any witness sees an item, and evaluated on held out twins. The one composition that caught the unanimous error did so because the same witness, asked to extract rather than decide, marked the decisive fact as undetermined.



1. Archive continuity and anti-repetition verification

Before designing any Round 41 experiment we rebuilt the archive ledger from the live research registry (prior-rounds-ledger.json, 34 indexed rounds). A keyword scan of registry titles, questions, and open questions for the Round 41 constructs (selection bias, disagreement mining, isomorphic, twin holdout, screening twin, hard pool, regression to the mean, consensus error, in sample) returned one hit: "hard pool" in the Round 40 handoff, which this round answers. The registry indexes titles and questions only, so we also compared the closest rounds by hand.

Prior round	Focus	How Round 41 differs
NCT-40 (Round 40)	Six composition topologies compared on 33 items	Round 41 tests the procedure for building a harder pool; the topology comparison is not repeated
Round 32	A blinded mutation holdout for a release gate	The holdout here guards a selection procedure against selection on noise, not a gate against overfitting
Round 29	Provenance twins: same content, different provenance	Twins here share a rule trace and verdict and differ in parties, dates, amounts, and record order
Round 14	Label uncertainty	Labels are not varied; the label contestable item F07 is disclosed, and external label review stays with the proposed Round 44 item

No packet, exhibit, or prompt task is reused from an earlier round as a test of the same question. The direct and exhibit prompts are the Round 40 prompts, unchanged, so that results are comparable. The twenty exhibits come from the candidate pool assembled in Round 39 and had never been put to a witness.

2. Where this round sits in the program

The published roadmap names Rounds 41 to 50 "Theory v1": an intended convergence into a first Theory of Collaborative Intelligence, keeping only hypotheses that survive testing. Round 40 proposed five propositions (P1 to P5) and a path for Rounds 41 to 50. This round tests the instrument that the path depends on: a pool that can separate compositions. It adds one proposition (P6) and revises the method for Round 42 and later rounds.

3. Design

3.1 Topologies and scoring

The topology codes follow Round 40. T1 is Witness A alone. T2 is the plurality of four direct verdicts with ties broken by Witness A. The 3-of-4 quorum releases a verdict only if three witnesses agree, otherwise it routes to review. T3 is one extraction by Witness C resolved by deterministic code that checks every quote against the source and fails closed. T6 runs extractions by Witness C and Witness B through the same resolver and releases only if both resolve to the same determinate verdict. The role pipeline and critique loop (T4, T5) are not run; rewriting versus flagging validators is the proposed Round 43 item.

A decision is a correct release, an unsafe release (a determinate verdict that differs from truth, including any release when truth is HOLD), a correct hold, or a false hold (a hold when a determinate verdict is supportable).

3.2 Isomorphic twins

Each authored family is one rule program rendered twice. The build computes each expected verdict from structured facts by the family's rule function and asserts that both twins produce the identical rule

trace and verdict. Twins differ in parties, dates, amounts, wording of records, and, for F10, the order in which records appear. The screening twin is S; the evaluation twin is E. Structured facts are never shown to witnesses.

3.3 Selection rule

A family is selected if the four direct verdicts on its screening twin are not unanimous. The pre-stated plan was to compare topologies on selected families in sample (on S) and on the held out twin (E), and to compare selected against unselected families.

3.4 A new extraction schema for authored families

Round 40's authored schema was built for conflicting records. Round 41 families depend on counts, thresholds, overrides, and definitions, so extraction uses a general condition ledger: each condition carries an exact condition quote, a status of MET, UNMET, or UNDETERMINED, and an exact evidence quote, plus the verdict if all are met and the verdict if any is unmet. The resolver fails closed to HOLD on any unverifiable quote, missing status, or missing mapping; any UNDETERMINED condition yields HOLD; otherwise any UNMET condition yields the unmet verdict. The real exhibit extraction prompt and resolver are the Round 40 ones, unchanged.

4. Related work

Benchmarks built by keeping items that current models fail on can be hard for those models while overstating difficulty for others: adversarial collection [S36], adversarial filtering of dataset artifacts [S37], and dynamic human in the loop collection [S38] all select items with a model in the loop. Fresh test sets drawn by the same procedure can show accuracy drops that are not explained by overfitting alone [S39], and minimally changed contrast sets probe whether a decision follows the rule rather than the surface [S40]; our twins are a contrast set held constant in rule and varied in surface. Panels and votes help when errors are independent [S09, S26, S35]; latent class models estimate error rates under conditional independence [S28], which is exactly what a consensus error violates. Abstention in language models is surveyed in [S24].

5. Hypotheses (stated before the live run)

- H41.1 Items selected for disagreement show lower accuracy in sample than on held out twins, for every topology.
- H41.2 Selection for disagreement inflates the measured gap between topologies.
- H41.3 Families authored to be hard by construction produce screening disagreement on at least some families, so topologies separate on accuracy.
- H41.4 Each witness gives the same verdict on both twins of a family.

- H41.5 Disagreement selection misses consensus errors: some errors fall on items where all witnesses agree.

6. Evidence category 1: deterministic diagnostic (E41-A)

6.1 Model

Items are easy (60 percent), hard (28 percent), or trap (12 percent). On hard items a witness's errors split evenly over the two wrong verdicts; on trap items they all fall on one wrong verdict. Error rates are fixed per witness and type (for example Witness A: 0.02, 0.30, 0.55), conditionally independent given type, and ordered as in the Round 40 measurements. A twin keeps its original's type with probability ρ and otherwise draws a type from the prior. All 81 answer vectors per type are enumerated exactly; nothing is sampled.

6.2 Results

Pool	T1 wrong release	T2 wrong release	Quorum wrong release	T2 minus T1 accuracy
Full pool	0.162	0.139	0.075	0.023
Selected, in sample	0.339	0.285	0.133	0.055
Selected, twin holdout, $\rho = 1.0$	0.303	0.274	0.147	0.028
Selected, twin holdout, $\rho = 0.7$	0.260	0.234	0.126	0.027
Selected, twin holdout, $\rho = 0.4$	0.218	0.193	0.104	0.025

Selection keeps 42.2 percent of items and moves the type mix from 60 / 28 / 12 to 20 / 56 / 24 (easy, hard, trap). Three structural facts follow:

1. **In sample error is inflated.** Even with perfect twins ($\rho = 1.0$), single witness and forced majority error rates are higher in sample than on the twin, because part of what selection selected was noise on the screening copy.
2. **The topology gap is inflated.** The forced majority's advantage over one witness is 0.055 in sample and 0.028 on the twin. A topology comparison run on the selected items themselves would roughly double the advantage it reports.
3. **Consensus errors are invisible to selection.** 13.4 percent of all forced majority errors, and all unanimous wrong answers (1.9 percent of the pool), are on items that are never selected. These are the errors a quorum releases. This is consistent with the quorum's wrong release rate rising from 0.133 in sample to 0.147 on the twin: selection removed its characteristic error from the sample.

6.3 Twin and window certificates

All 10 families produce identical rule traces and verdicts on both twins. No twin pair shares an amount except one figure in F08, and the only shared dates are fiscal quarter ends required by the rule (F07,

F08). All 20 exhibit windows were built by one deterministic rule (the first 1,800 characters and 4,600 characters starting 600 before an anchored effectiveness passage); every anchor matched; no window contains a dollar sign before a digit.

7. Evidence category 2: live authored twins (E41-B)

7.1 Protocol

All calls ran at temperature 0. Witness A ran at maximum reasoning; Witness B, Witness C, and Witness D ran at standard reasoning. Screening: 4 direct calls per S twin. Evaluation: 4 direct calls plus 2 extractions per E twin. All 100 authored calls returned and parsed.

7.2 Families

Family	Construction	Truth
F01	Deemed consent if no objection within ten Business Days; a scheduled holiday and an objection sent after 5:00 p.m. on the last day make the objection late	PASS
F02	A Default waived by lenders holding exactly fifty percent when Required Lenders means more than fifty percent	FAIL
F03	An extension needs a fee and one of two guaranty deliveries "delivered after the date of this Second Amendment"; the only delivery predates it by one day	FAIL
F04	Agent approval given by a successor after acceptance; the resigned agent objects later	PASS
F05	A success fee clause is satisfied, but a notwithstanding clause bars any fee for a Closing during a Suspension Period	FAIL
F06	A cure pays fee and interest but not the five percent late charge, which the license makes immediately due	FAIL
F07	A margin step down uses the most recent quarter whose statements "have been delivered"; later statements exist, would reverse the outcome, and have no delivery record	HOLD
F08	A "greater of" basket, where the percentage prong governs and a prior fiscal year investment must be excluded	PASS
F09	A rate cut needs an officer's certificate covering every property on Schedule 4, and Schedule 4 is missing	HOLD
F10	Authority comes from the most recent incumbency certificate; an older certificate omits the signer	PASS

7.3 Results

Measure	Result
Families with any screening disagreement	0 / 10
Families with any evaluation disagreement	0 / 10
Direct accuracy per witness, S and E	9 / 10 and 9 / 10 for each of Witness A to Witness D
Witness twin pairs with the same verdict	40 / 40
Unanimous wrong items	F07S, F07E

Topology on E twins	Correct release	Unsafe	Correct hold	False hold	Calls per decision
T1 single witness	8	1 (F07E)	1	0	1
T2 forced majority	8	1 (F07E)	1	0	4
3-of-4 quorum	8	1 (F07E)	1	0	4
T3 extract then resolve	8	1 (F07E)	1	0	1
T6 disagreement routed	7	0	2	1 (F03E)	2

The screening twins gave the same topology results as the evaluation twins. Because no family was selected, the planned in sample versus holdout comparison on selected families has no live items. H41.1 and H41.2 therefore rest on E41-A alone.

7.4 Failure walk-through: F07

The term loan sets the margin from "the most recent fiscal quarter for which financial statements have been delivered to Lender under Section 6.1 before that interest period begins". The June quarter statements were delivered with a receipt; the September quarter statements exist, were approved before the interest period, would push the ratio above the threshold, and "the file contains no transmittal or receipt record". The authored label is HOLD: the outcome turns on a delivery the packet neither shows nor rules out.

All four witnesses released PASS on both twins. Witness B's rationale on the screening twin states the gap and then resolves it the wrong way: "The packet does not establish delivery of the September 30 statements before November 16. The most recent delivered statements reported a Leverage Ratio of 2.88, making the 2.75 percent margin correct." The other witnesses wrote that the September statements "were not delivered". Absence of a record was read as proof of absence.

In T6, Witness C as extractor marked the margin condition MET from the June statements, so T3 released. Witness B as extractor, on the same packet where it had answered PASS directly, marked two conditions UNDETERMINED and wrote: "If June controls, 2.41 is 0.09 below 2.50 and 3.5 percent applies, while if September controls, 2.63 is 0.13 above 2.50 and 4.0 percent applies." The resolver held. The same witness, moved from deciding to extracting, decorrelated from its own direct answer.

F07's label is contestable: an auditor could adopt a rule that undocumented deliveries are treated as not made. No external reviewer checked the label before the run. We report it, and the result without it: excluding F07, every topology except T6 is 9 / 9 and T6 is 8 / 9 with one false hold.

7.5 The T6 false hold: F03E

Witness B's extraction was substantively right (fee condition MET, guaranty condition UNMET) but its evidence quote joined two separate records with a line break into one string. The span check correctly could not find that string in the packet, and the resolver failed closed. The routed hybrid paid for its safety on F07 with a hold caused by quote formatting, not by a reading error.

8. Evidence category 3: real SEC EDGAR exhibits (E41-c)

8.1 Track

Twenty public exhibits, none put to a witness in any earlier round, were fixed before any call, each with its accession number and archive link, under the Round 39 and 40 label rule: CERTIFY only if the exhibit text alone establishes the principal operative effect; receipt or delivery of counterparts, fees, certificates, and closings elsewhere are conditions the exhibit does not establish. Strata: H1 "effective as of" language with conditions (10), H2 execution and counterparts only (3), H3 signature deemed to confirm satisfaction (2), H4 recitals stating other conditions are satisfied (1), and H5 instruments effective on their face beside condition language that does not govern them (4 , CERTIFY). All 20 exhibits were evaluated with the same prompts, witnesses, and settings: four direct calls and one extraction call by Witness C per exhibit. All 100 exhibit calls returned and parsed.

ID	Filer	Exhibit	Accession	Stratum	Expected
Y01	Stellus Private Credit BDC	EX-10.01	0001104659-25-097942	H1	HOLD
Y02	Ares Strategic Income Fund	EX-10.1	0001104659-25-119310	H1	HOLD
Y03	MasTec, Inc.	EX-10.2	0001193125-26-297552	H1	HOLD
Y04	Tiptree Inc.	EX-10.1	0001393726-26-000034	H1	HOLD
Y05	Novanta Inc.	EX-10.1	0001193125-26-262867	H1	HOLD
Y06	CoreCivic, Inc.	EX-10.1	0001193125-25-304483	H1	HOLD
Y07	Kelly Services, Inc.	EX-10.4	0001193125-26-031768	H1	HOLD
Y08	Werner Enterprises, Inc.	EX-10.1	0000793074-26-000120	H1	HOLD
Y09	Eton Pharmaceuticals, Inc.	EX-10.1	0001437749-26-012381	H1	HOLD
Y10	FTC Solar, Inc.	EX-10.1	0001193125-26-120582	H1	HOLD
Y11	Cenntro Inc.	EX-10.1	0001140361-25-020044	H2	HOLD
Y12	Coca-Cola Consolidated, Inc.	EX-10.3	0001628280-25-050682	H2	HOLD
Y13	BRC Group Holdings, Inc.	EX-10.1	0001213900-26-023145	H2	HOLD
Y14	Piedmont Realty Trust, Inc.	EX-10.2	0001042776-25-000146	H3	HOLD
Y15	Quest Resource Holding Corp	EX-10.2	0001193805-25-000429	H3	HOLD
Y16	PennantPark Floating Rate Capital Ltd.	EX-10.1	0001193125-26-082897	H4	HOLD
Y17	Sonic Automotive, Inc.	EX-10.2	0001628280-25-000245	H5	CERTIFY
Y18	Avnet, Inc.	EX-10.2	0000008858-25-000021	H5	CERTIFY
Y19	Applied Digital Corp.	EX-10.5	0001493152-25-006268	H5	CERTIFY
Y20	Granite Construction Inc.	EX-10.2	0000861459-25-000037	H5	CERTIFY

8.2 Results

Topology	Correct release	Unsafe	Correct hold	False hold
T1 single witness	4	2 (Y11 , Y13)	14	0

Topology	Correct release	Unsafe	Correct hold	False hold
T2 forced majority	4	2 (Y11 , Y13)	14	0
3-of-4 quorum	4	0	16	0
T3 extract then resolve	4	0	16	0

Without the three execution only exhibits (Y11 , Y12 , Y13), all four topologies are 17 / 17. Y11 and Y13 are the only exhibits with any disagreement, so the disagreement mined subset of the real track has two items, both in one stratum, and on both the split was between the same two pairs of witnesses.

8.3 Failure walk-through: Y13 and Y11

The amendment "shall become effective when it shall have been executed by the Administrative Agent and when the Administrative Agent shall have received counterparts hereof that, when taken together, bear the signatures of each of the other parties hereto". The filed copy carries conformed signatures. Witness A and Witness C certified, reasoning that the signatures in the exhibit satisfy the condition "within the four corners of the document". Witness B and Witness D held: "The exhibit shows signatures but does not itself establish the required receipt." The 2-2 tie went to Witness A under T2, so the forced majority released. The quorum could not reach three and routed to review; Witness C's extraction listed the counterpart clause as a condition and the resolver held.

Y11 repeats the pattern. The amendment "shall become effective as of the date first written above ... upon satisfaction of each of the following conditions precedent", and the only listed condition is that it "shall have been duly executed and delivered by each party." Witness A certified because "the filed exhibit is itself the executed instrument"; Witness C certified for the same reason. Witness B and Witness D held because the window "does not establish that every party executed and delivered the amendment." T2 released on the tie; the quorum routed to review, Witness C's extraction listed the execution and delivery clause as a condition, and the resolver held.

This is the same stratum that Round 40 flagged, and its label depends on a rule (receipt is a condition the filing does not establish) that a reader may reasonably reject; hence the results with and without it.

9. Findings and interpretations

9.1 Measured findings

1. **Deterministic.** Disagreement selection doubles the measured forced majority advantage over one witness (0.055 in sample against 0.023 on the full pool and 0.028 on a perfect twin).
2. **Deterministic.** 13.4 percent of forced majority errors lie on items that are never selected, including every unanimous error.
3. **Live.** 0 / 10 hard by construction families produced any screening disagreement; the selection step selected nothing.
4. **Live.** Twin stability was 40 / 40 witness twin pairs.

5. **Live.** Both authored errors were unanimous (F07S , F07E). Only T6 held F07E , with one false hold (F03E) caused by a joined quote.
6. **Live.** On 20 real exhibits, two disagreements occurred (Y11 and Y13 , both execution only, both split Witness A and Witness C against Witness B and Witness D). The forced majority released both on a tie; the quorum and the resolver held both. All other 18 exhibits were unanimous and correct.

9.2 Interpretation

The Round 40 handoff assumed that harder items would surface as disagreement. On this panel they do not: when the witnesses err, they err together. A procedure that keeps only split items would have built a pool of zero authored items and two real ones from a single stratum, and it would have discarded F07 , the one item on which every composition except T6 released unsafely. E41-A shows the same pattern in exact form. Selection on disagreement finds independent mistakes and overstates how much composition helps against them, while correlated mistakes, which no vote can fix, are invisible to it.

Two practical consequences follow. First, a pool that separates compositions has to be hard by construction, with labels fixed before any witness sees it, and should be evaluated on a held out twin. Second, the most useful source of decorrelation observed here was not a different witness but a different role: the same witness that released F07 directly marked it undetermined when asked to extract. That moves the Round 42 decorrelation question from "which witness pairs" toward "which role and prompt pairs".

9.3 Hypothesis status

Hypothesis	Evidence	Status
H41.1	E41-A only: in sample accuracy is below twin accuracy for T1 , T2 , and the quorum at every rho; no live selected items	Supported (structural only)
H41.2	E41-A only: gap 0.055 in sample against 0.028 on twin	Supported (structural only)
H41.3	0 / 10 families with screening disagreement	Rejected
H41.4	40 / 40 witness twin pairs stable	Supported
H41.5	E41-A 13.4 percent; live F07 unanimous on both twins	Supported

10. Toward Theory v1

Round 40 named propositions P1 to P5. Round 41 adds one and revises one.

- **P6. Disagreement is a biased detector.** Disagreement among witnesses detects decorrelated error, not difficulty. Pools selected by disagreement overstate the advantage of composition and omit consensus errors. Evidence: E41-A ; F07 . Removal test: a panel on which disagreement mined pools and constructed pools give the same topology ranking on held out twins.
- **P4 revised. Decorrelation is the safety budget, and role is a lever.** Evidence added: Witness B direct PASS and extraction UNDETERMINED on F07E . Removal test: measured joint error across the

same witness in two roles, against two witnesses in one role (proposed for Round 42).

11. Method changes for later rounds

1. Build test pools by construction, with labels computed from structured facts and fixed before any witness call.
 2. Write each item as isomorphic twins; select or tune on one twin and report on the other.
 3. Treat unanimous items as the place to look for consensus errors, not as solved items.
 4. In the span check, allow an evidence field to cite more than one record as separate quotes, so that formatting does not cause holds.
-

12. Synthesis review (interpretation, not evidence)

After all results were recorded, Witness A at maximum reasoning received the structural and live results and was asked to separate what the data show from what they suggest, and to list threats to validity. Its review is archived verbatim (`synthesis-review.json`). It agreed that the live authored round gave no test of the screen then select step because the witnesses were perfectly correlated, judged F07 more likely a shared handling choice about an undetermined condition than independent noise, and warned that T6 trading an unsafe release for a false hold reflects a hold on undetermined policy rather than better discrimination. It also noted that the exhibit track skews toward HOLD. Its proposed next check on undetermined condition families with both HOLD and PASS labels is adopted in Section 14. The review was written before the full exhibit track was complete and is archived unchanged; the completed track adds one disagreement (Y11) of the same kind it discussed. The review is a second reading, not a source of any number in this paper.

13. Limitations and threats to validity

- The pools are small (10 families, 20 exhibits). Every live difference between topologies is one or two items; no ranking is statistically supported.
- The exhibit track is weighted toward HOLD (16 of 20), which favors holding topologies.
- Families and labels were authored by the research team. F07 and the execution only stratum have plausible alternative labels. No external reviewer checked labels before the run.
- Twins share rule traces by construction, so twin stability may reflect surface similarity as well as reasoning.
- E41-A uses assumed type priors and error rates; it shows the direction and mechanism of selection effects, not their size on any real panel.
- Temperature 0 runs were made once per call; topologies share calls, so their errors are not independent samples.

14. Next check

Author a constructed pool in which undetermined fact items like F07 appear with both HOLD and PASS labels, then measure, on held out twins, the joint error of the same witness in direct and extraction roles against two different witnesses in the same role.

15. References

- [S09] Xuezhi Wang et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171, 2022.
- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S26] Pat Verga, Sebastian Hofstatter, et al. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796, 2024.
- [S28] A. P. Dawid, A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Applied Statistics 28(1), 1979. doi:10.2307/2346806.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. <https://efits.sec.gov/LATEST/search-index> and <https://www.sec.gov/Archives/edgar/data/>.
- [S35] Krishna K. Ladha. The Condorcet Jury Theorem, Free Speech, and Correlated Votes. American Journal of Political Science 36(3), 1992.
- [S36] Yixin Nie, Adina Williams, Emily Dinan, et al. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv:1910.14599, 2019.
- [S37] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, et al. Adversarial Filters of Dataset Biases. arXiv:2002.04108, 2020.
- [S38] Douwe Kiela, Max Bartolo, Yixin Nie, et al. Dynabench: Rethinking Benchmarking in NLP. arXiv:2104.14337, 2021.
- [S39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? arXiv:1902.10811, 2019.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.